Comparative transcriptomics of primary cells in vertebrates

Comparative transcriptomics of primary cells in vertebrates

Tanvir Alam[1], Saumya Agrawal[2], Jessica Severin[2], Robert S. Young[3,4], Robin Andersson[5], Erik Arner[2], Akira Hasegawa[2], Marina Lizio[2], Jordan Ramilowski[2], Imad Abugessaisa[2], Yuri Ishizu[6], Shohei Noma[2], Hiroshi Tarui[7], Martin S. Taylor[4], Timo Lassmann[2,8], Masayoshi Itoh[9], Takeya Kasukawa[2], Hideya Kawaji[2,9], Luigi Marchionni[10], Guojun Sheng[11], Alistair Forrest[2,12], Levon M. Khachigian[13], Yoshihide Hayashizaki[9], Piero Carninci[2], Michiel de Hoon[2]

[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar;
[2]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan;
[3]Centre for Global Health Research, Usher Institute, University of Edinburgh, United Kingdom;
[4]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, United Kingdom;
[5]The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark;
[6]RIKEN Center for Brain Science, Wako, Japan;
[7]RIKEN Center for Life Science Technologies, Yokohama, Japan;
[8]Telethon Kids Institute, University of Western Australia, Subiaco, Western Australia, Australia;
[9]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Japan;
[10]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA;
[11]International Research Center for Medical Sciences (IRCMS), Kumamoto University, Kumamoto, Japan;
[12]Harry Perkins Institute of Medical Research, and the Centre for Medical Research, University of Western Australia, QEII Medical Centre, Perth, Western Australia, Australia;
[13]Vascular Biology and Translational Research, School of Medical Sciences, Faculty of Medicine, University of New South Wales, Sydney, Australia.

Abstract

Gene expression profiles in homologous tissues have been observed to be different between species, which may be due to differences between species in the gene expression program in each cell type, but may also reflect differences between species in cell type composition of each tissue. Here, we compare expression profiles in matching primary cells in human, mouse, rat, dog, and chicken using Cap Analysis Gene Expression (CAGE) and short RNA (sRNA) sequencing data from FANTOM5. While we find that expression profiles of orthologous genes in different species are highly correlated across cell types, many genes were differentially expressed between species within homologous cell types. Expression of genes with products involved in transcription, RNA processing, and transcriptional regulation was more likely to be conserved, while that of genes encoding proteins involved in intercellular communication were more likely to have diverged during evolution. Conservation of expression correlated positively with the evolutionary age of genes, suggesting that divergence in expression levels of genes critical for cell function was restricted during evolution. Motif activity analysis showed that both promoters and enhancers are activated by the same transcription factors in different species. An analysis of expression levels of mature miRNAs and of primary miRNAs identified by CAGE revealed that evolutionary old miRNAs are more likely to have conserved expression patterns than young miRNAs. We conclude that key aspects of the regulatory network are conserved, while differential expression of genes involved in cell-to-cell communication may contribute greatly to phenotypic differences between species.

Introduction

Vertebrate organisms consist of 100’s of cell types, with more than 400 cell types defined in human (Vickaryous and Hall 2006). Traditionally, cell types have been defined by their tissue of

origin as well as by their cellular phenotypes including morphology, staining properties, enzyme histochemistry, and cell surface marker recognition by antibodies (Vickaryous and Hall 2006). Cell type characterization has been supplemented by molecular approaches such as molecular fingerprinting (Arendt 2008) as well as genome-wide profiling of the transcriptome of primary cells (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). To this end, the Human Cell Atlas initiative aims to comprehensively define human cell types by performing transcriptome analysis in single cells on a massive scale (Regev et al. 2017).

Evolution of anatomy is thought to primarily depend on the evolution of gene expression patterns and regulation, rather than the evolution of the encoded protein sequences (King and Wilson 1975; Britten and Davidson 1971). While comparative studies have shown that gene expression programs in matching tissues are largely conserved between species (Chan et al. 2009; Brawand et al. 2011; Merkin et al. 2012; Su et al. 2002), many genes were found to be differentially expressed (Su et al. 2002, Lin et al. 2014; Yue et al. 2014). Although such expression differences between human and mouse for specific genes may be due in part to differences in cell type composition of the analyzed tissues (Breschi et al. 2017), little overlap was found in terms of differentially expressed genes between human and mouse in dynamic studies of primary cells during erythropoiesis (Pishesha et al. 2014), and of primary macrophages upon stimulation by lipopolysaccharide (Schroder et al. 2012) or by glucocorticoid (Jubb et al. 2016). Collectively, these findings suggest that also in matching primary cells many genes are differentially expressed between species. As cells with an identical cellular phenotype may display distinct and disparate molecular phenotypes, the question what key transcriptomic features define a cell type is raised (Arendt et al. 2016).

The confounding effects of cell type composition in tissue-based studies can be avoided by comparing the transcriptome of different species in homologous primary cells. Here, we present

a comparative analysis of genome-wide expression in vertebrate species profiled in FANTOM5 (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014; Lizio, Mukarram, et al. 2017; Lizio, Deviatiiarov, et al. 2017) to elucidate patterns of gene expression conservation during evolution.

Results

The FANTOM5 collection contains Cap Analysis Gene Expression (CAGE) data for 3 primary cell types in human, mouse, rat, dog, and chicken, and for an additional 12 cell types in human and mouse only (Supplemental Table S1). We identified 15,538, 14,915, 13,759, and 8,696 protein-coding genes in mouse, rat, dog, and chicken, respectively, with a one-to-one orthologous gene in human, and 6,561 protein-coding genes with one-to-one orthologs in all five species (see Methods for details). Principal Component Analysis (PCA) of all human and mouse samples revealed a liver-specific cluster, a mesenchymal cluster, and a hematopoietic cluster (Fig. 1A), and similarly PCA for cell types with CAGE data available in all five species showed a hepatocyte cluster and a mesenchymal cluster (Fig. 1B). Within each cluster, samples tended to cluster by species (Fig. 1), consistent with the "species signal" phenomenon observed previously (Musser & Wagner 2015).

Expression levels of pairs of orthologous genes were positively correlated across cell types, with median Pearson's correlation values ranging from 0.38 to 0.72 ($p < 10^{-100}$, mouse, rat, and dog; $p = 2.2 \times 10^{-42}$, chicken; Fig. 2A,B). Nevertheless, in specific cell types we found significant differences in absolute expression of orthologs in different species (Fig. 2A,C). Pairwise differential expression analysis between genes in human and their orthologs in mouse, rat, dog, or chicken for each primary cell type in FANTOM5 revealed that on average 52% of expressed genes were differentially expressed (Benjamini-Hochberg corrected $p < 0.1$) between the two species (Fig. 2C and Supplemental Table S2).

In each species, we defined the dominant promoter for each gene as the most highly expressed promoter associated with the gene. The genomic region of the dominant promoter of more than 80% of genes in mouse, rat, and dog, and 50% of genes in chicken had an orthologous region in the human genome; the majority of those overlapped the corresponding human dominant promoter (Fig. 3A). Genes were more likely to be differentially expressed if their dominant promoter was located in a genomic region that did not have an orthologous genome sequence in the human genome (Fisher combined $p < 10^{-100}$; Fig. 3B and Supplemental Fig. S1), suggesting that gain or loss of promoter sequence regions during evolution contributes to the emergence of gene expression differences between species.

We hypothesized that genes critical for cellular functioning would both be more conserved and have more conserved expression patterns, and indeed found the expression levels of evolutionarily older genes to be more conserved (Fisher combined $p < 10^{-100}$; Fig. 4A and Supplemental Fig. S2). Gene Ontology analysis of differentially expressed genes showed that genes with products involved in transcription, RNA processing, and transcriptional regulation, were more likely to have conserved expression levels, whereas genes encoding proteins localized to the plasma membrane and extracellular space as well as signaling proteins were

most likely to be differentially expressed (Fig. 4B and Supplemental Table S3). This suggests that the transcriptional program in each cell tends to be conserved during evolution, while genes in the periphery of the transcriptional regulatory network, especially those involved in cellular communication, tend to diverge in expression.

As an independent confirmation, we applied integrative correlation analysis (Parmigiani et al. 2004) by first calculating the correlations across cell types between all genes for human and mouse separately, and then the correlation across orthologous genes between corresponding rows in these two correlation matrices. This yielded the correlation-of-correlations, or integrative correlation coefficient, as a measure of the degree of expression conservation during evolution for each gene. We then ranked genes based on their integrative correlation coefficient, and performed gene set enrichment analysis to identify biological processes most conserved or most divergent between the two species (see Methods section).

The integrative correlation coefficient values ranged between -0.52 to 0.59, and their observed distribution was skewed to the right, with a median of 0.25 (Supplemental Fig. S3A and Supplemental Table S4), suggesting that, overall, gene expression profiles tend to be conserved between human and mouse. Similarly to our conclusions for Gene Ontology analysis of differentially expressed genes, fundamental cellular processes involved in cell homeostasis and maintenance tended to rank higher in integrative correlation analysis, while gene sets encompassing processes associated with cell-to-cell signaling and other biological processes taking place in the extra-cellular space (e.g., neuronal and synapse development) were more likely to rank lower, suggesting less conserved underlying networks (Supplemental Fig. S3B and Supplemental Table S4).

As a complement to the differential gene expression analysis, we calculated the correlation across genes for each cell type and species. Expression levels were positively correlated within each species as well as between species for related cell types (Supplemental Fig. S4, S5), suggesting the relative ranking of genes by their expression level tends to be conserved. The correlation value decreased exponentially as a function of phylogenetic distance between species, and dropped off most rapidly for mesenchymal stem cells compared to aortic smooth muscle cells and hepatocytes (Supplemental Fig. S6). Consistent with the differential gene expression results, expression levels were more highly correlated for genes for which the dominant promoter had an orthologous genome region in human compared to genes for which the dominant promoter did not have an orthologous genome region (Fisher combined $p < 10^{-100}$; Supplemental Fig. S7 and S8), as well as for evolutionarily ancient genes compared to recent genes (Fisher combined $p < 10^{-100}$; Supplemental Fig. S9 and S10). A Gene Ontology analysis of correlation values again showed that genes with functional roles associated with RNA biology in the nucleus tended to have conserved expression levels, while genes with functions associated with the plasma membrane, extracellular space, and signaling had lower correlation values (Supplemental Fig. S11 and Supplemental Table S3).

To confirm these findings in an independent gene expression data set, we performed differential expression analysis on previously published RNA-seq expression data for endometrial stromal fibroblast primary cells in human, rat, rabbit, ferret, cow, and opossum (Kin et al. 2016). We again found that evolutionarily ancient genes were more likely to have conserved expression levels compared to recent genes (Fisher combined $p = 1.0 \times 10^{-11}$; Supplemental Fig. S12). Results of Gene Ontology analysis of differentially expressed genes for these data were highly consistent with those observed in the FANTOM5 samples (Supplemental Fig. S12), including evidence of rapid evolution of signaling pathways as observed previously (Kin et al., 2016). A comparative analysis of RNA-seq expression data in matching tissues in human and mouse

(The ENCODE Project Consortium 2012) also showed preferential conservation of expression levels of evolutionarily ancient genes (Supplemental Fig. S13A,B), and yielded similar patterns of Gene Ontology enrichment (Supplemental Fig. S13C).

To understand how evolution of the transcriptional regulatory network affects evolution of gene expression, we used the MotEvo sequence motif analysis software (Arnold et al. 2012) for the 190 motifs compiled in SwissRegulon (Pachkov et al. 2013) to identify potential transcription factor binding sites (TFBSs) in the human, mouse, rat, dog, and chicken genomes. We evaluated the TFBS prediction accuracy using ChIP-seq data (Supplemental Table S5) for transcription factors associated with each motif (Supplemental Fig. S14). Conservation between species of the expression patterns of orthologous genes depended on the concordance in TFBS presence in the promoter of each gene (Supplemental Fig. S15), demonstrating the contribution of cis-regulatory evolution to expression divergence between species. To analyze trans-regulatory evolution, we performed motif activity analysis (Suzuki et al. 2009), which uses linear decomposition of genome-wide gene expression patterns based on the TFBSs found in the promoter of each gene, resulting in motif activities representing the average expression level of genes with a predicted binding site for each motif. Fig. 5 shows the broadly expressed transcription factor TP53 (Fig. 5A), the hematopoietic lineage-specific RUNX transcription factors (Fig. 5B), and the motif associated with the hepatocyte-specific HNF4A transcription factor (Fig. 5C) as examples of motifs with activities highly correlated between human and mouse. In contrast, the motif associated with the testis-specific transcription factor SPZ1 did not show evidence of activation either in human or mouse, as testis was not included in our samples (Fig. 5D). In general, motif activities were highly correlated across samples between human and mouse ($p = 5.5 \times 10^{-25}$, Mann-Whitney U test), rat ($p = 3.9 \times 10^{-9}$), dog ($p = 4.5 \times 10^{-6}$), and chicken ($p = 9.2 \times 10^{-4}$), compared to randomized pairs of motifs (Fig. 5E and Supplemental Table S6).

We then asked if enhancers likewise were activated by the same transcription factors in different species. Enhancers were previously identified in human and mouse from FANTOM5 CAGE data by searching for a characteristic bidirectional expression pattern (Andersson et al. 2014). We predicted enhancers in rat, dog, and chicken by applying the same pipeline on the FANTOM5 CAGE data in these species (Supplemental Table S7), and used the CAGE expression level at each enhancer as a measure of its activity (Andersson et al. 2014). For each species, the motif activity calculated from gene promoter expression profiles correlated with the motif activity based on enhancer expression profiles (human, $p = 1.2 \times 10^{-20}$, Mann-Whitney *U* test), mouse ($p = 5.6 \times 10^{-22}$), rat ($p = 5.6 \times 10^{-5}$), dog ($p = 2.5 \times 10^{-5}$), and chicken ($p = 5.7 \times 10^{-4}$) (Supplemental Fig. S16), indicating that in each species enhancers are activated by the same transcription factors as promoters. Between species, the motif activity calculated from enhancer expression profiles were highly correlated between human and mouse ($p = 1.6 \times 10^{-18}$, Mann-Whitney *U* test), rat ($p = 2.6 \times 10^{-6}$), dog ($p = 0.0032$), and chicken ($p = 0.044$) (Fig. 5F and Supplemental Table S6). We conclude that both promoters and enhancers are activated by the same transcription factors in different species.

Next, we extended our comparative analysis to the expression levels of microRNAs (miRNAs). miRNAs are small non-coding RNA (typically 22 nts) that silence mRNA post-transcriptionally and regulate biological processes such as cell growth and differentiation by functional effects on direct targets and regulatory networks (Bracken et al. 2016). In the FANTOM5 collection, short RNA (sRNA) sequencing data for matching primary cell types in different species were available for aortic smooth muscle cells (Supplemental Table S1 and Supplemental Table S8). We annotated known (Supplemental Table S9) and candidate novel (Supplemental Table S10) miRNAs in rat, dog, and chicken in the same way as done previously (De Rie et al. 2017) for human and mouse. Differential expression analysis between human and mouse, rat, dog or

chicken showed that about half of the orthologous miRNAs had statistically significant different expression levels in the two species (Fig. 6A and Supplemental Table S11). Dividing miRNAs into three categories based on their evolutionary age revealed that evolutionarily older miRNAs were more likely to have conserved expression levels than younger miRNAs (Fisher combined $p = 1.2 \times 10^{-4}$) (Fig. 6C).

Previously, we showed that CAGE data can be used to reliably infer the promoter of the primary miRNA (pri-miRNA) transcript, and that the corresponding CAGE expression levels can be used as a proxy for the expression level of the mature miRNA (De Rie et al. 2017). We manually curated pri-miRNA promoters previously identified computationally for mouse (De Rie et al. 2017), and, using the same approach, identified pri-miRNA promoters for miRNAs in rat, dog, and chicken (Supplemental Table S12). In aortic smooth muscle cells, expression levels of the mature miRNA measured by sRNA sequencing correlated with the CAGE expression level of the pri-miRNA for mouse, rat, dog, and chicken (Supplemental Fig. S17). The curated primary miRNA promoter annotations as well as expression levels of the mature and primary miRNA are visualized and available for download through an interactive web interface at https://fantom.gsc.riken.jp/zenbu/reports/#FANTOM_miRNA_atlas.

Using these promoters together with previously curated pri-miRNA promoters for human (De Rie et al. 2017), we performed differential expression analysis of miRNAs in human compared to mouse, rat, dog, and chicken. In aortic smooth muscle cells in mouse, rat, dog, and chicken, log-ratios of mature miRNA expression levels, as measured by sRNA sequencing, correlated well with the log-ratios for pri-miRNAs, as measured by CAGE expression data (Supplemental Fig. S18), and among the miRNAs differentially expressed in both data sets, more than 80% showed concordant up- or down-regulation of the mature miRNA and the pri-miRNA, suggesting

that few of the identified differentially expressed miRNAs were false positives (Supplemental Fig. S18).

Differential CAGE expression analysis of pri-miRNAs revealed that the majority of expressed orthologous miRNAs have different expression levels in human compared to mouse, rat, dog, and chicken (Fig. 6B and Supplemental Table S13), consistent with the results obtained for mature miRNAs (Fig. 6A). We found significantly fewer differentially expressed miRNAs for evolutionarily old miRNAs compared to evolutionarily recent miRNAs for 12 out of 24 pairwise comparisons, a further 7 showed the same pattern without reaching statistical significance, 5 showed an opposite pattern without reaching statistical significance, and 0 showed a statistically significant opposite pattern (Fisher combined $p = 4 \times 10^{-12}$; Fig. 6D). Therefore, using CAGE as a proxy for miRNA expression allowed us to demonstrate that the patterns observed for mature miRNAs by sRNA sequencing for a single cell type (Fig. 6C) can be found across a wide variety of cell types.

Discussion

Comparative studies have shown considerable differences in the gene expression levels in matching tissues of different species (Su et al. 2002; Lin et al. 2014; Yue et al. 2014), which at least in part is due to differences in tissue composition between species (Breschi et al. 2017). However, our analysis reveals that this cannot be the sole explanation, as considerable expression level differences are also observed between matching primary cell types, indicating that the same cellular phenotype associated with traditionally defined cell types can be achieved by widely different molecular networks.

Our findings suggest that expression levels of regulators tend to be conserved across species, while genes peripheral in the regulatory network, especially those involved in cellular communication, are more likely to have divergent expression patterns. Previously reported examples include the terminal differentiation of erythroid precursors from early to late erythroblasts, where the same transcriptional regulators and other proteins important for erythropoiesis were induced or repressed in human and mouse, suggesting that the core regulatory program of erythroid differentiation remained conserved (Pishesha et al. 2014). In contrast, genes regulated during development showed a different response between human and mouse (Pishesha et al. 2014), indicating that the response of genes to the regulators of erythropoiesis had evolved since the evolutionary split of human and mouse. Similarly, comparing lipopolysaccharide-stimulated macrophages between human and mouse showed enriched differences in the transcriptome of genes encoding proteins involved in cellular communication such as cell surface receptors, inflammatory cytokines, chemokines, and their intracellular signaling pathways (Schroder et al. 2012). Phenotypic differences between species

at the organismal level may thus be primarily due to differences in the interaction between cells (Ramilowski et al. 2015).

Orthologous transcription factors typically recognize the same DNA sequence motif in human and mouse (Cheng et al. 2014), as changes in the consensus motif during evolution would simultaneously affect a large number of genes and may be too disruptive. By the same argument, we can expect that expression levels of transcriptional regulators to be conserved between species. As a salient example of the conservation of regulatory programs, we previously found that human enhancer sequences could be activated by orthologous transcription factors in corresponding tissues in human and zebrafish (Andersson et al. 2014). In contrast, genomic binding sites of conserved transcription factors have diverged extensively between human and mouse (Odom et al. 2007), suggesting a rewiring of the peripheral regulatory network during evolution.

Due to their modular nature, enhancer regulatory elements are particularly amenable to rewiring, as their cell type- and state-specific usage (Andersson et al. 2014) allow changes in their regulatory connections in specific conditions while avoiding pleiotropic deleterious effects on the organism in general (Carroll 2008). For example, differences in the transcriptome response of human and mouse primary macrophages stimulated by glucocorticoid were previously found to be associated with the turnover of glucocorticoid receptor binding sites at enhancers (Jubb et al. 2016). Similarly, the cell type- and state-specific usage of different promoters associated with a gene (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014) avoids the constraints placed by pleiotropy and allows gain and loss of promoters to contribute significantly to the evolution of species (Young et al. 2015). In our analysis, we indeed find that changes in gene expression levels are associated with the gain and loss of promoter sequence regions during evolution.

Our analysis further shows that the conservation of regulatory programs is not limited to transcriptional regulation but extends to miRNAs. Our comparative analysis of miRNA expression revealed that older miRNAs are more likely to have conserved expression levels than more recent miRNAs, suggesting that highly conserved miRNAs have stronger evolutionary constraints on their expression levels. As an example, we found conservation across human, mouse, rat, dog, and chicken of pri-miRNA expression levels in aortic smooth muscle cells of miR-22, which modulates a range of target genes including *MECP2*, *HDAC4* and *MECOM* and is a key regulator of smooth muscle cell phenotype switching and neointima formation (Yang et al. 2018).

The Human Cell Atlas aims to create a comprehensive map of cell types in the human body by profiling gene expression levels in single cells from healthy tissues (Regev et al. 2017). Our comparative analysis suggests that differences in the regulatory signature (Arendt et al. 2016), rather than the overall gene expression patterns, are the key requirement for distinguishing cell types.

Methods

Genome assembly version

For consistency with previous FANTOM5 publications (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014; Lizio, Mukarram, et al. 2017; Lizio, Deviatiiarov, et al. 2017), we used genome assemblies hg19 (human), mm9 (mouse), rn6 (rat), canFam3 (dog), and galGal5 (chicken) for our analysis. Previously it was shown that 99.75% and 99.94% of CAGE peaks in human and mouse, respectively, could be converted unambiguously to the

recent genome versions hg38 (human) and mm10 (mouse), with an expression correlation value larger than 0.99 both for human and mouse (Abugessaisa et al., 2017). For 231 (human) and 202 (mouse) miRNAs included in the comparative analysis shown in Supplemental Table S13, the genomic distance between each pre-miRNA and the corresponding pri-miRNA promoter (Supplemental Table S12) was identical between genome assembly versions for 216 (human) and 199 (mouse) miRNAs, and differed by less than 10 base pairs for 227 (human) and 202 (mouse) miRNAs, suggesting that the genome assembly version used had minimal effects on the analysis results.

Identification of orthologous genes

For each gene in mouse, rat, dog, and chicken defined in Ensembl (Zerbino et al. 2018) release 85, we retrieved the orthologous human gene, if defined, in an 'ortholog_one2one' relationship with it in the Ensembl Compara multi-species database (Vilella et al. 2009). This yielded 16,217 (human-mouse), 15,486 (human-rat), 15,861 (human-dog), 11,950 (human-chicken), and 10,237 (in all five species) pairs of orthologous genes, of which 15,893 (human-mouse), 15,207 (human-rat), 15,482 (human-dog), 11,873 (human-chicken), and 10,208 (in all five species) were protein-coding. Using the most recent Ensembl release available for each genome assembly (release 75 for human genome assembly hg19, release 67 for mouse genome assembly mm9, release 85 for rat genome assembly rn6 and dog genome assembly canFam3, and release 92 for chicken genome assembly galGal5), we obtained the transcription start site for all transcripts associated with each gene, defined a ±500 base pair promoter region around each transcription start site, and merged overlapping regions. Genes for which any of the associated regions had more than 10% unidentified nucleotides (N) in their genome sequence were removed from the analysis. The number of remaining orthologous protein-coding genes

was 15,538 (human-mouse), 14,915 (human-rat), 13,759 (human-dog), 8,696 (human-chicken), and 6,561 (in all five species).

Gene expression analysis

Gene expression quantitation is described in detail in the Supplemental Methods. Differential gene expression analysis was performed on the raw counts using DESeq2 (Love et al. 2014) version 1.22.1 with a threshold of 0.1 on the Benjamini-Hochberg adjusted *p*-value. PCA as well as all correlation calculations (except in Integrative Correlation Coefficient analysis, described below) were performed on variance-stabilized gene expression data generated as follows. First, we used DESeq2 (Love et al. 2014) version 1.22.1 to estimate, for each cell type in each species, the asymptotic dispersion of expression counts between replicates, and then calculated its average value α across cell types and species. Next, we calculated the total tag count for each sample, divided these totals by their median across samples to obtain the normalization factors, and divided the counts of each sample by the corresponding factor to obtain normalized count data x. We then applied the variance-stabilizing transformation (Anders & Huber, 2010) to the normalized count data x:

$$x' = \frac{2\,\mathrm{arcsinh}\left(\sqrt{\alpha x}\right) - \log\alpha - \log 4}{\log 2}$$

The variance-stabilized gene expression data x' were averaged across replicates for each cell type and for each species.

For each pairwise comparison in Fig. 2B, we calculated Pearson's correlation across cell types between each pair of orthologous genes. Next, we randomly permuted the gene pairings, calculated the correlation across cell types to find the background distribution, and performed the Mann-Whitney *U* test comparing the set of correlation values for pairs of orthologous genes to the set of correlation values for randomly permuted pairs. We also calculated a background

distribution for pairs of orthologous genes after permuting the samples, as well as the cumulative distribution of correlation values for an uncorrelated bivariate normal distribution. For the pairwise comparisons shown in Supplemental Fig. S4–S11, we calculated Pearson's correlation between the two species for each cell type across orthologous genes. For Supplemental Fig. S15, we calculated Pearson's correlation between the two species for each pair of orthologous genes across cell types.

Promoter conservation analysis

Orthologous genomic regions of promoters across species were identified by applying liftOver (Hinrichs et al. 2006) on chain files downloaded from the University of California, Santa Cruz website (http://hgdownload.cse.ucsc.edu/downloads.html).

Gene conservation analysis

For each gene in human, we identified the HomoloGene group of homologous genes to which it belonged in release 68 of the NCBI HomoloGene database (NCBI Resource Coordinators 2018). If the HomoloGene group included mammals only or vertebrates only, then the gene was classified as restricted to mammals or restricted to vertebrates, respectively. Alternatively, the gene was classified as conserved in bilateria if the HomoloGene group included bilateria in non-vertebrate lineages. To assess the statistical significance of the increase or decrease in conservation of expression in the three classes, the bilaterian, vertebrate, and mammalian class were represented by an equidistant indicator variable, and applied the maximum likelihood method to fit a linear regression model under the Poisson distribution to the number of differentially expressed genes in each class. The corresponding *p*-value was calculated using the likelihood-ratio test. The overall *p*-value was calculated by combining the *p*-values for the pairwise comparisons using Fisher’s method.

Gene Ontology analysis

Gene Ontology annotations were downloaded on June 10, 2018 from the GOA database (Huntley et al. 2015). Statistical significance of over- or underrepresentation of a Gene Ontology term among differentially expressed genes was calculated using Fisher's exact test, where an expression-matched background was created by selecting the 10 closest genes in expression in human for each differentially expressed gene. The overall *p*-value was calculated by combining the *p*-values for the pairwise comparisons using Fisher's method.

RNA-seq expression data analysis

Accession numbers for ENCODE (The ENCODE Project Consortium 2012) and endometrial stromal fibroblast (Kin et al. 2016) RNA-seq gene expression data are provided in the Supplemental Methods. Gene conservation and Gene Ontology analysis of these data sets were performed as described above.

Integrative Correlation Coefficient analysis

Integrative Correlation Coefficient analysis (Parmigiani et al. 2004) ranks genes based on the degree to which their expression profiles are comparable between datasets.

For human and mouse separately, we constructed a CAGE expression matrix (normalized to t.p.m.) for the 15,538 genes in common between human and mouse, averaging biological replicates by taking the median, and performed quantile normalization separately for each expression matrix. Next, we calculated the correlation between each pair of genes, again for human and mouse separately, across cell types to obtain one correlation matrix for human and one correlation matrix for mouse. We then calculated Pearson's correlation between human and mouse for corresponding rows in these two correlation matrices to obtain the correlation-of-

correlations, or integrative correlation coefficient, for each gene. The null distribution was obtained by randomly permuting samples 10,000 times, as described previously (Parmigiani et al. 2004), using MergeMaid (Cope et al. 2004) version 2.56.0. Analysis of Functional Annotation (AFA) (Kortenhorst et al. 2013, Marchionni et al. 2017, Ross et al. 2011) was conducted by performing a one-sided Wilcoxon rank-sum test to compare the integrative correlation coefficient values of genes in each cellular component (CC) and biological process (BP) Gene Ontology category (extracted using the "org.Hs.eg.db" R/Bioconductor package version 3.8.2), requiring at least 10 genes, to those of remaining genes, using the Benjamini-Hochberg multiple testing correction method. All analyses were performed using the R/Bioconductor "RTopper" package (version 1.30.0) (Tyekucheva et al. 2011).

Multiple genome alignment, TFBS prediction, and motif activity analysis

The 100-way multiple genome alignment of human genome assembly hg19 against 99 vertebrate species and the 30-way multiple genome alignment of the mouse genome assembly mm9 against 29 vertebrate species were downloaded from the University of California, Santa Cruz website (http://hgdownload.cse.ucsc.edu/downloads.html), the species in the 30-way mouse alignment being a subset of the species in the 100-way human alignment. For the same set of 30 species, we performed pairwise genome alignments of the rat, dog, and chicken genome against each of the 29 remaining species for the genome assemblies listed in Supplemental Table S14 (see Supplemental Methods for details). Pairwise alignments were merged into a multiple genome alignment using multiz (Blanchette et al. 2004) version 11.2 using the phylogenetic tree of the 30 species was extracted from the 191-way phylogenetic tree in 191way.nh distributed as part of the UCSC Genome Browser bioinformatics utilities (Kuhn et al. 2013) release 366 (June 5, 2018). Genome-wide TFBS predictions and motif activity analysis

were performed as described previously (Arner et al. 2015), with minor modifications as described in the Supplemental Methods. The multiple genome alignment files, genome-wide locations and scores of predicted TFBSs, and motif activity scripts are available at http://fantom.gsc.riken.jp/5/suppl/Alam_et_al_2020/; motif activity scripts are also included in the Supplemental Code.

Enhancer identification

The previously calculated set of permissive enhancers (Arner et al. 2015) was used for human (65,423 enhancers) and mouse (44,459 enhancers). For rat, dog, and chicken, we first created a mask for all ±500 base pair windows around the 5' end of transcripts in the NCBI Entrez Gene database (Brown et al. 2015), downloaded on 13 November 2017, as well as all windows within 200 base pairs of exons defined in the same database. We then applied the bidir_enhancers script (Andersson et al. 2014) to all FANTOM5 CAGE libraries in rat, dog (Lizio, Mukarram, et al. 2017), and chicken (Lizio, Deviatiiarov, et al. 2017) using the calculated mask, resulting in 9,372 (rat), 10,649 (dog), and 44,625 (chicken) enhancers.

MicroRNA analysis

Short RNA libraries were produced, sequenced, and processed as described previously (De Rie et al. 2017) using the same RNA samples as used for CAGE expression profiling (Lizio, Deviatiiarov, et al. 2017; Lizio, Mukarram, et al. 2017). Short RNA libraries not described previously are listed with their matching CAGE library in Supplemental Table S1. Annotation of miRNAs, candidate novel miRNA prediction, and miRNA promoter identification were performed as described in the Supplemental Methods. Orthologous miRNAs were identified by performing global alignment of mature miRNA sequences between species, followed by manual curation.

The evolutionary age of miRNAs was established based on the set of species in which miRNAs of each family were annotated in miRBase release 21 (Kozomara and Griffiths-Jones 2014).

Data access

All raw and processed sequencing data generated in this study have been submitted to the DNA Data Bank of Japan (DDBJ; https://www.ddbj.nig.ac.jp/) under accession number DRA008211. All custom scripts generated in this study are available as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

This work was supported by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to Y.H.; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT to Y.H.; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant from MEXT to the RIKEN Center for Integrative Medical Sciences; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to Y.H.; NIH-NCI award R01CA200859 and the JSPS Fellowship S19058 to L.M. We gratefully acknowledge the computational resources of the HOKUSAI supercomputer system provided by RIKEN under project number Q18305/Q19305, which enabled us to perform the pairwise genome alignments as well as the genome-wide TFBS predictions by MotEvo.

Author contributions:

T.A., L.M., and M.d.H. analyzed the data with the help of J.R., R.A., E.A., R.S.Y., M.S.T., T.L., A.H., H.K.; G.S. provided RNA samples; Y.I., S.N., H.T., and M.I. produced the sequencing libraries; I.A., M.L., H.K., and T.K. managed the data; S.A. curated the primary miRNA annotations; J.S. created the interactive web interface for miRNA expression visualization; A.R.R.F. and M.d.H. designed the study; T.A., L.M., and M.d.H. wrote the manuscript with the help of L.M.K. and P.C.; P.C. and Y.H. supervised the FANTOM5 project.

References

Abugessaisa I, Noguchi S, Hasegawa A, Harshbarger J, Kondo A, Lizio M, Severin J, Carninci P, Kawaji H, Kasukawa T. 2017. FANTOM5 CAGE Profiles of Human and Mouse Reprocessed for GRCh38 and GRCm38 Genome Assemblies. *Scientific Data* **4**: 170107.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet* **9**: 868–882.

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD, Wagner GP. 2016. The origin and evolution of cell types. *Nat Rev Genet* **17**: 744–757.

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**: 1010–1014.

Arnold P, Erb I, Pachkov M, Molina N, Van Nimwegen, E. 2012. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* **28**: 487–494.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–715.

Bracken CP, Scott HS, Goodall GJ. 2016. A network-biology perspective of microRNA function

and dysfunction in cancer. *Nat Rev Genet* **17**: 719–732.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Breschi A, Gingeras TR, Guigó, R. 2017. Comparative transcriptomics in human and mouse. *Nat Rev Genet* **18**: 425–440.

Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* **46**: 111–138.

Brown GR, Hem V, Katz KS, et al. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* **43**(Database issue): D36–D42.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**: 25–36.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, et al. 2009. Conservation of core gene expression in vertebrate tissues. *J Biol* **8**: 33.

Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515**: 371–375.

Cope L, Zhong X, Garrett E, Parmigiani G. 2004. MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat Appl Genet Mol Biol* **3**: Article29.

De Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, Åström G, Babina M, Bertin N, Burroughs AM, et al. 2017. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat Biotechnol* **35**: 872–878.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT).2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**(Database issue): D590–D598.

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. 2015. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res* **43**(Database issue): D1057-D1063.

Jubb AW, Young RS, Hume DA, Bickmore WA. 2016. Enhancer turnover Is associated with a divergent transcriptional response to glucocorticoid in mouse and human macrophages. *J Immunol* **196**: 813–822.

Kin K, Maziarz J, Chavan AR, Kamat M, Vasudevan S, Birt A, Emera D, Lynch VJ, Ott TL, Pavlicev M, Wagner GP. 2016. The transcriptomic evolution of mammalian pregnancy: Gene expression innovations in endometrial stromal fibroblasts. *Genome Biol Evol* **8**: 2459–2473.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Kortenhorst MS, Wissing MD, Rodríguez R, Kachhap SK, Jans JJ, Van der Groep P, Verheul HM, Gupta A, Aiyetan PO, van der Wall E, et al. 2013. Analysis of the genomic response of human prostate cancer cells to histone deacetylase inhibitors. *Epigenetics* **8**: 907–920.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**(Database issue): D68-D73.

Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161.

Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman

WC, et al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A* **111**: 17224–17229.

Lizio M, Deviatiiarov R, Nagai H, Galan L, Arner E, Itoh M, Lassmann T, Kasukawa T, Hasegawa A, Ros MA, et al. 2017. Systematic analysis of transcription start sites in avian development. *PLoS Biol* **15**: e2002887.

Lizio M, Mukarram AK, Ohno M, Watanabe S, Itoh M, Hasegawa A, Lassmann T, Severin J, Harshbarger J, Abugessaisa I, et al. 2017. Monitoring transcription initiation activities in rat and dog. *Sci Data* **4**: 170173.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Marchionni L, Hayashi M, Guida E, Ooki A, Munari E, Jabboure FJ, Dinalankara W, Raza A, Netto GJ, Hoque MO, Argani P. 2017. MicroRNA expression profiling of Xp11 renal cell carcinoma. *Hum Pathol* **67**: 18–29.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**: 1593–1599.

Musser JM, Wagner GP. 2015. Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called "species signal". *J Exp Zool B Mol Dev Evol* **324**: 588–604.

NCBI Resource Coordinators 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**(D1): D8–D13.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**: 730–732.

Pachkov M, Balwierz PJ, Arnold P, Ozonov E, Van Nimwegen E. 2013. SwissRegulon, a

database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res* **41**(Database issue): D214–D220.

Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E. 2004. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res* **10**: 2922–2927.

Pishesha N, Thiru P, Shi J, Eng JC, Sankaran VG, Lodish HF. 2014. Transcriptional divergence and conservation of human and mouse erythropoiesis. *Proc Natl Acad Sci U S A* **111**: 4103–4108.

Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, Itoh M, Kawaji H, Carninci P, Rost B, Forrest AR. 2015. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* **6**: 7866.

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. 2017. The human cell atlas. *eLife* **6**: e27041.

Ross AE, Marchionni L, Phillips TM, Miller RM, Hurley PJ, Simons BW, Salmasi AH, Schaeffer AJ, Gearhart JP, Schaeffer EM. 2011. Molecular effects of genistein on male urethral development. *J Urol* **185**: 1894–1898.

Schroder K, Irvine KM, Taylor MS, Bokil NJ, Le Cao KA, Masterman KA, Labzin LI, Semple CA, Kapetanovic R, Fairbairn L, et al. 2012. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc Natl Acad Sci U S A* **109**: E944-E953.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**: 4465–4470.

Suzuki H, Forrest ARR, Van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T,

Ravasi T, Hasegawa Y, De Hoon MJL, et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**: 553–562.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Tyekucheva S, Marchionni L, Karchin R, Parmigiani G. 2011. Integrating diverse genomic data using gene sets. *Genome Biol* **12**: R105.

Vickaryous MK, Hall BK. 2006. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc* **81**: 425–455.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**: 327–335.

Yang F, Chen Q, He S, Yang M, Maguire EM, An W, Afzal TA, Luong LA, Zhang L, Xiao Q. 2018. miR-22 is a novel mediator of vascular smooth muscle cell phenotypic modulation and neointima formation. *Circulation* **137**: 1824–1841.

Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, FANTOM Consortium, Bickmore WA, et al. 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res* **25**: 1546–1557.

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**: 355–364.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**(D1): D754–D761.

Figure captions

Figure 1. Gene expression PCA. (*A*) PCA for all samples of cell types in common between human and mouse. (*B*) PCA for all samples of cell types in common between all five species.

Figure 2. Differential gene expression analysis. (*A*) Expression profile of *HNF4A*, *ELF2*, and *FOXO1* as examples of genes with an expression profile highly correlated across cell types between species, but with significant expression level differences between species in specific cell types. (*B*) Cumulative distribution of Pearson's correlation *r* across cell types in gene expression between human and mouse, rat, dog, or chicken. The number *N* of expressed orthologous genes included in the distribution is shown in the vertical axis label, and the estimated median value of *r* is indicated on the horizontal axis of each graph. The background distribution of *r* obtained by randomizing genes (solid) or randomizing samples (dashed) as well as the theoretical background distribution of *r* for an uncorrelated bivariate normal distribution (dotted) are shown in grey. The statistical significance was calculated using the Mann-Whitney *U* test comparing Pearson's correlation values for orthologs to the background distribution of *r* for randomly paired genes between human and mouse, rat, dog, or chicken. Note that the median correlation values are not directly comparable between species, as the sets of orthologous genes are different. (*C*) Differential gene expression analysis of orthologous genes in human compared to mouse, rat, dog, and chicken. The red and blue bars correspond to the percentage of expressed orthologous genes with significantly (Benjamini-Hochberg corrected $p < 0.1$) higher and lower expression, respectively, in human compared to mouse, rat, dog, or chicken. The number *N* of orthologous genes expressed in each cell type is shown on the right.

Figure 3. Promoter analysis of differentially expressed genes. (*A*) Percentage of genes in mouse, rat, dog, and chicken for which the dominant promoter was located in a genome region that had an orthologous genome region in human, and the percentage that the orthologous region contained the dominant promoter for the orthologous gene in human. (*B*) Percentage of differentially expressed genes in each cell type depending on whether the genomic region of the dominant promoter in each species had an orthologous genomic region in the human genome. The one-sided *p*-value calculated using Fisher's exact test is shown on the right, together with the number *N* of expressed genes in each cell type.

Figure 4. Conservation and Gene Ontology analysis of differentially expressed genes. (*A*) Percentage of differentially expressed genes in each cell type as a function of age of the most recent common ancestor. The one-sided *p*-value of a Poisson regression model against the evolutionary age category is shown on the right, together with the number *N* of expressed genes in each cell type with an annotation in the NCBI HomoloGene database. (*B*) Gene Ontology analysis of differentially expressed genes. The *p*-value, calculated using Fisher's exact test, of overrepresentation or underrepresentation of differentially expressed genes in each Gene Ontology category compared to an expression-matched set of background genes is shown in red and blue, respectively.

Figure 5. Motif Activity analysis. (*A*-*D*) Examples of calculated motif activities in human and mouse for motifs associated with the broadly expressed transcription factor TP53 (*A*), the hematopoietic lineage-specific RUNX transcription factors (*B*), the hepatocyte-specific HNF4A transcription factor (*C*), and the testis-specific transcription factor SPZ1 (*D*). Each of the 15 matching cell types between human and mouse is shown as a dot. The blood cell types $CD19^+$ B cells, $CD4^+$ T cells, $CD8^+$ T cells, common myeloid progenitors, and granulocyte macrophage progenitors are shown in red for the RUNX motif, and the liver cell types hepatic sinusoidal

endothelial cells, hepatic stellate cells (lipocytes), and hepatocytes are shown in green for the motif associated with HNF4A. (*E*-*F*) Cumulative distribution of Pearson's correlation *r* across cell types in motif activity for promoters (*E*) and enhancers (*F*) between human and mouse, rat, dog, and chicken. The estimated median value of *r* is indicated on the horizontal axis of each graph. As a background distribution, we calculated the same correlation between pairs of different motifs in human and mouse, rat, dog, and chicken. The Mann-Whitney *U* test *p*-value comparing the actual correlation values to the correlation values of the background distribution is shown for each comparison.

Figure 6. Differential miRNA expression analysis. (*A*) Differential expression analysis of miRNAs using FANTOM5 sRNA sequencing data in aortic smooth muscle cells in human compared to mouse, rat, dog or chicken. The red and blue bars correspond to the percentage of expressed orthologous miRNAs with significantly (Benjamini-Hochberg corrected $p < 0.1$) higher and lower expression, respectively, in human compared to mouse, rat, dog, or chicken. The number *N* of expressed orthologous miRNAs in each comparison is shown on the right. (*B*) Differential expression analysis of miRNAs in human compared to mouse, rat, dog, and chicken; using CAGE expression of the pri-miRNA as a proxy for the expression level of the mature miRNA. The red and blue bars correspond to the percentage of expressed orthologous miRNAs with significantly (Benjamini-Hochberg corrected $p < 0.1$) higher and lower expression, respectively, in human compared to mouse, rat, dog, or chicken. The number *N* of expressed orthologous miRNAs in each comparison is shown on the right. (*C*) Percentage of miRNAs differentially expressed in each comparison, separately based on the evolutionary age of each miRNA. The one-sided *p*-value of a Poisson regression model against the evolutionary age category is shown on the right, together with the number *N* of expressed orthologous miRNAs in each comparison. (*D*) Percentage of miRNAs differentially expressed in each comparison, separately based on the evolutionary age of each miRNA; using CAGE expression of the pri-

miRNA as a proxy for the expression level of the mature miRNA. The one-sided *p*-value of a Poisson regression model against the evolutionary age category is shown on the right, together with the number *N* of expressed orthologous miRNAs in each comparison.

Supplemental Methods

Gene expression quantitation

For each promoter region for each gene, we found the total CAGE tag count, summing over all libraries included in this study. For each gene, the region with the highest total CAGE tag count was identified as the dominant promoter, and the position with the highest CAGE tag count within this region as the representative transcription start site of the gene. For each gene, the CAGE expression level in a CAGE library was defined as the sum of CAGE tags across all regions associated with the gene. CAGE expression data were normalized to tags-per-million (t.p.m.) by dividing by the total sum of CAGE tags associated with genes, and multiplying by 1,000,000. Genes with an expression level of at least 10 t.p.m. were included in the analysis.

RNA-seq expression data

ENCODE (The ENCODE Project Consortium 2012) RNA-seq gene expression data were downloaded from the ENCODE portal (Davis et al. 2018) (https://www.encodeproject.org/) as gene quantifications in tsv format with sample annotations in json format for the following accession numbers: ENCFF005MLW, ENCFF051UVH, ENCFF084WWG, ENCFF115BZB, ENCFF122LQH, ENCFF134RPA, ENCFF143IYG, ENCFF152DMV, ENCFF162KBI, ENCFF201BIE, ENCFF203RCU, ENCFF227RBV, ENCFF280HQA, ENCFF362GHJ, ENCFF375LAX, ENCFF421SLK, ENCFF428HNN, ENCFF442XPO, ENCFF456GZP, ENCFF485EYP, ENCFF507RNA, ENCFF508QDF, ENCFF509LQW, ENCFF517ZTC, ENCFF547TAC, ENCFF549OGG, ENCFF603OKL, ENCFF625HJC, ENCFF637YVK, ENCFF664JKG, ENCFF677ZIY, ENCFF686JQP, ENCFF750HMK, ENCFF780DAF, ENCFF784ZTQ, ENCFF798PSE, ENCFF804BIT, ENCFF804WTK, ENCFF809RAX,

ENCFF850LMK, ENCFF866LBS, ENCFF878LJT, ENCFF878UHQ, ENCFF908GIP, ENCFF911YUO, ENCFF916CFV, ENCFF926FDN, ENCFF937HOV. RNA-seq gene expression count tables for endometrial stromal fibroblast (Kin et al. 2016), their normalized expression levels in tags per million, and gene ortholog associations were downloaded from the NCBI Gene Expression Omnibus under accession number GSE67659.

Pairwise genome alignments

Preprocessing of the genome sequences and postprocessing of the alignments was performed using partitionSequence.pl, blastz-normalizeLav, lavToPsl, axtChain, chainAntiRepeat, chainMergeSort, chainPreNet, chainNet, netSyntenic, netFilter, netToAxt, axtSort, and axtToMaf, which are part of the UCSC Genome Browser bioinformatics utilities (Kuhn et al. 2013) release 366 (June 5, 2018). First, for each pairwise alignment we used partitionSequence.pl to split the target and query genome sequences into segments of 20 Mb, with the target sequence segments overlapping by 10 kb. Sequences were aligned using lastz (Harris 2007) version 1.03.66 with parameters shown in Supplemental Table S15. Alignment coordinates were corrected using blastz-normalizeLav and converted to .psl format using lavToPsl. Alignments were chained using axtChain with parameters shown in Supplemental Table S15 and further processed using chainAntiRepeat, chainMergeSort, chainPreNet, chainNet, netSyntenic, and netFilter. The best alignments were extracted using netToAxt, followed by axtSort and axtToMaf to generate a .maf (multiple alignment format) file.

Genome-wide TFBS prediction

For human and mouse, alignments for human, macaque, mouse, rat, dog, horse, cow, opossum, and chicken were extracted from the 100-way (human) and 30-way (mouse, rat, dog, and chicken) multiple genome alignments; for rat, dog, and chicken, we used the 30-way

multiple genome alignments described in the main Methods section of this manuscript. Extracted alignments were improved as described previously (Arner et al. 2015) using the T-Coffee (Notredame et al. 2000) multiple sequence aligner version 9.01. We ran MotEvo (Arnold et al. 2012) version 1.01 on these alignments as described previously (Arner et al. 2015), but using the 190 motifs in the SwissRegulon (Pachkov et al. 2013) release of July 13, 2015, and with a background probability for the four nucleotides that was calculated by counting their frequency across the genome.

Motif activity analysis

The scripts make_profile.py, associate_tfbs.py, and calculate_motif_activity.py, available at http://fantom.gsc.riken.jp/5/suppl/Alam_et_al_2020/ and in the Supplemental Code, provide an implementation of the motif activity analysis methodology described previously (Suzuki et al. 2009). The density of predicted transcription factor binding sites with respect to the representative transcription start site was calculated for each motif was calculated using the script make_profile.py with default options. Using the calculated density, predicted binding sites were associated with the dominant promoter of each gene by the script associate_tfbs.py with default options. The density of predicted transcription factor binding sites in a ±500 base pair window around enhancers was calculated using the script make_profile.py with options --upstream=500 --downstream=500 --symmetric. Using the calculated density, predicted binding sites were associated with each enhancer by the script associate_tfbs.py with default options. We then selected motifs with at least 50 predicted binding sites both for promoters and in enhancers in each pair of species, resulting in 154 (human-mouse), 151 (human-rat), 149 (human-dog), and 94 (human-chicken) motifs. Motif activities were calculated for promoters and enhancers separately for the selected motifs using the script calculate_motif_activity.py with option ‘-n 0’.

MicroRNA analysis

We used release 21 of the miRBase database (Kozomara and Griffiths-Jones 2014), lifted over for rat and chicken to genome assembly rn6 and galGal5 respectively, as our reference set of known miRNAs; one pre-miRNA in rat and 26 pre-miRNAs in chicken could not be lifted over and were dropped. Pre-miRNAs were classified as robust or permissive (Supplemental Table S9) as described previously (De Rie et al. 2017).

Candidate novel miRNAs were identified using miRDeep2 (Friedländer et al. 2012), resulting in 229 (rat), 249 (dog), and 180 (chicken) predicted pre-miRNAs, including 169 (rat), 179 (dog), and 128 (chicken) known pre-miRNAs and 59 (rat), 74 (dog), and 55 (chicken) candidate novel pre-miRNAs (Supplemental Table S10).

Promoters for miRNAs in rat, dog, and chicken were identified using the same approach as applied previously for human and mouse (De Rie et al. 2017), using transcripts annotated for rat, dog, or chicken in the NCBI Entrez Gene database (Brown et al. 2015), downloaded on 13 November 2017, as candidate primary miRNAs. Identified promoters for all robust miRNAs in mouse, rat, dog, and chicken were curated manually by two annotators (Supplemental Table S12).

Supplemental References

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**(D1): D794–D801.

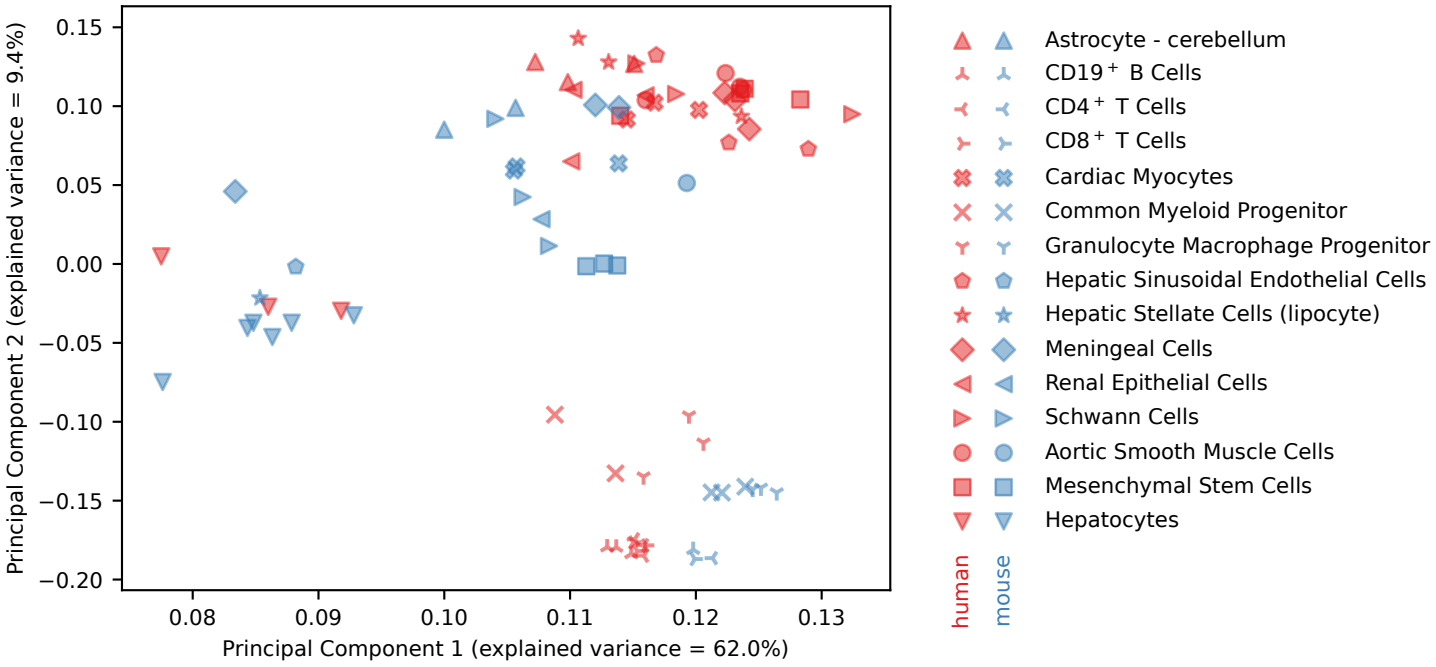
Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**: 37–52.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. Ph.D. thesis. Pennsylvania State University, State College, PA.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.

Figure 1

A


Principal Component 2 (explained variance = 9.4%)
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0.08
0.09
0.10
0.11
0.12
0.13
Principal Component 1 (explained variance = 62.0%)
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
human
mouse

B

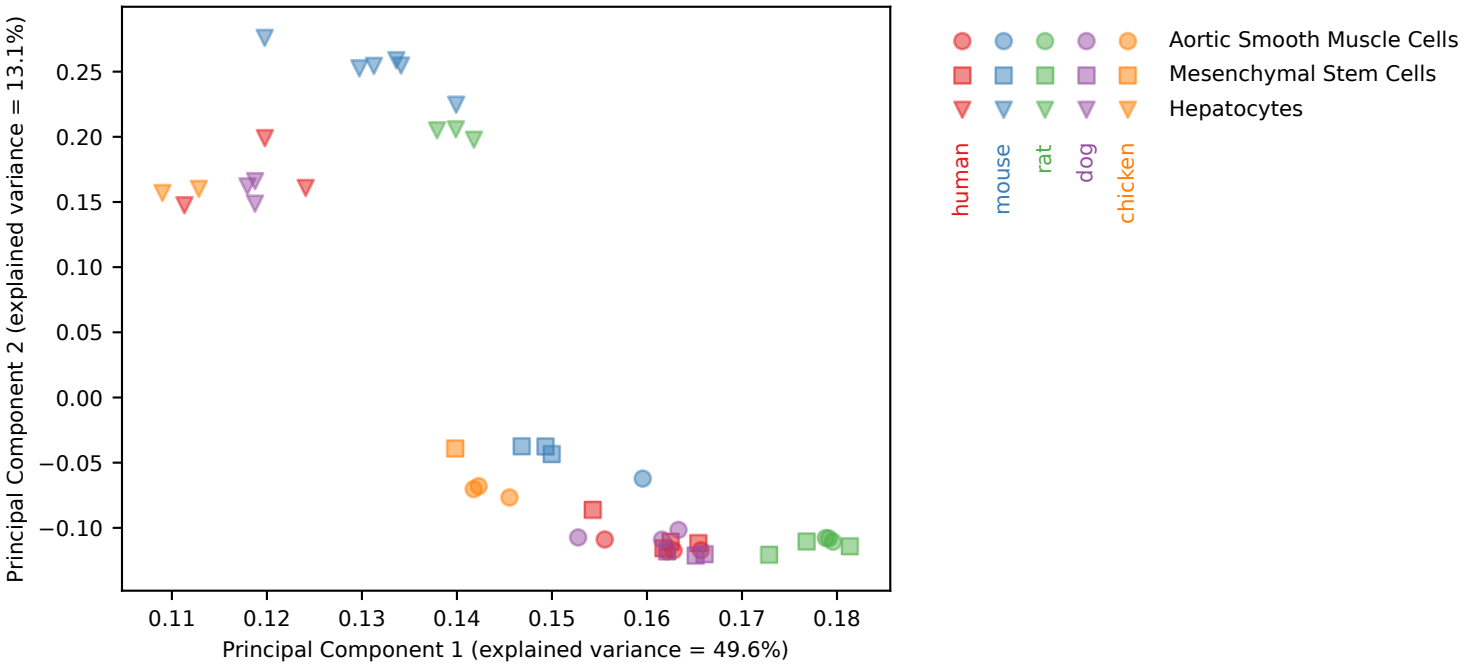

Principal Component 2 (explained variance = 13.1%)
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
Principal Component 1 (explained variance = 49.6%)
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
human
mouse
rat
dog
chicken

Figure 2

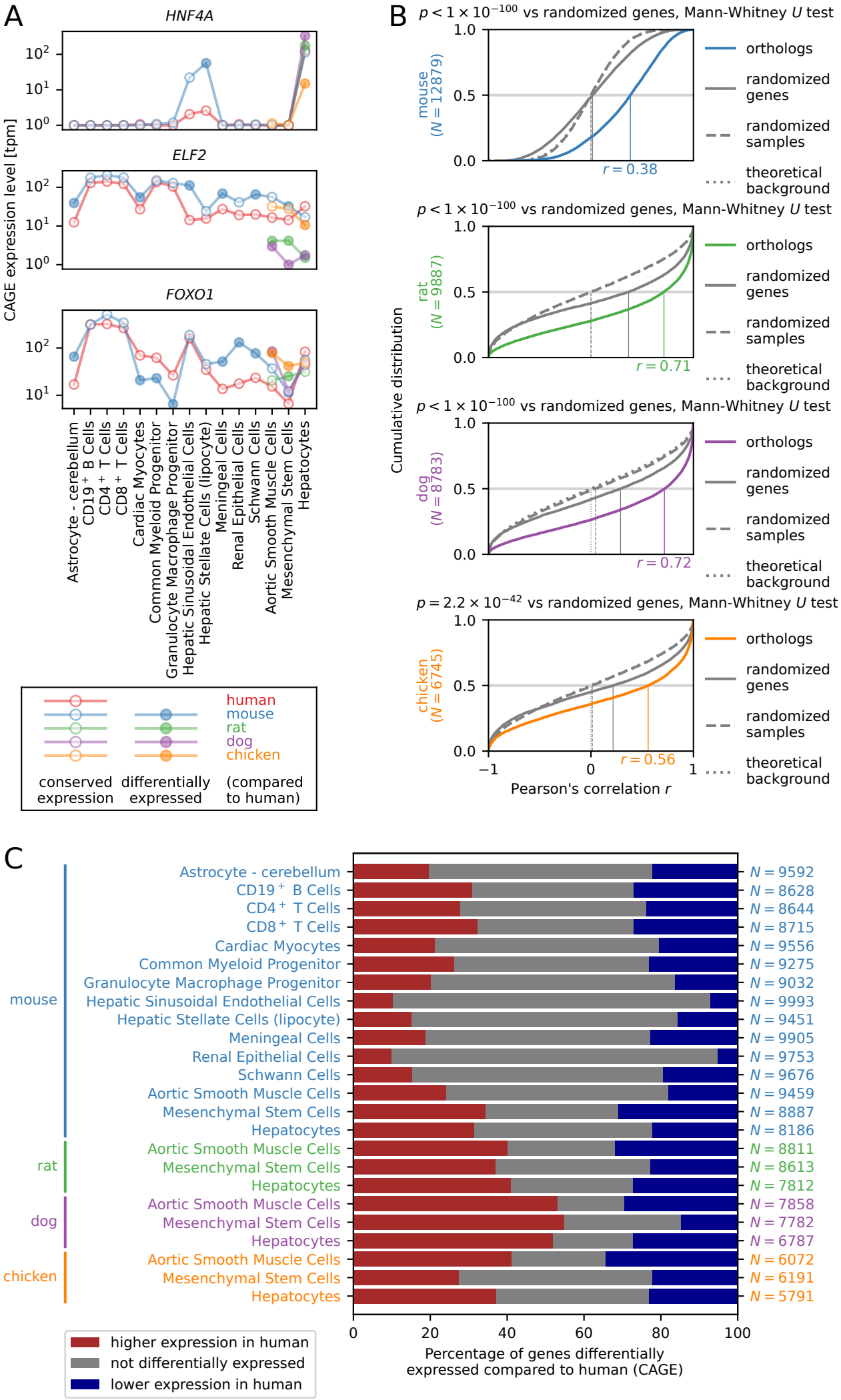

A
HNF4A
ELF2
FOXO1
CAGE expression level [tpm]
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
human
mouse
rat
dog
chicken
conserved expression
differentially expressed
(compared to human)
B
p < 1 × 10−100 vs randomized genes, Mann-Whitney U test
mouse (N = 12879)
r = 0.38
rat (N = 9887)
r = 0.71
dog (N = 8783)
r = 0.72
p = 2.2 × 10−42 vs randomized genes, Mann-Whitney U test
chicken (N = 6745)
r = 0.56
Cumulative distribution
Pearson's correlation r
orthologs
randomized genes
randomized samples
theoretical background
C
mouse
rat
dog
chicken
N = 9592
N = 8628
N = 8644
N = 8715
N = 9556
N = 9275
N = 9032
N = 9993
N = 9451
N = 9905
N = 9753
N = 9676
N = 9459
N = 8887
N = 8186
N = 8811
N = 8613
N = 7812
N = 7858
N = 7782
N = 6787
N = 6072
N = 6191
N = 5791
higher expression in human
not differentially expressed
lower expression in human
Percentage of genes differentially expressed compared to human (CAGE)

Figure 3

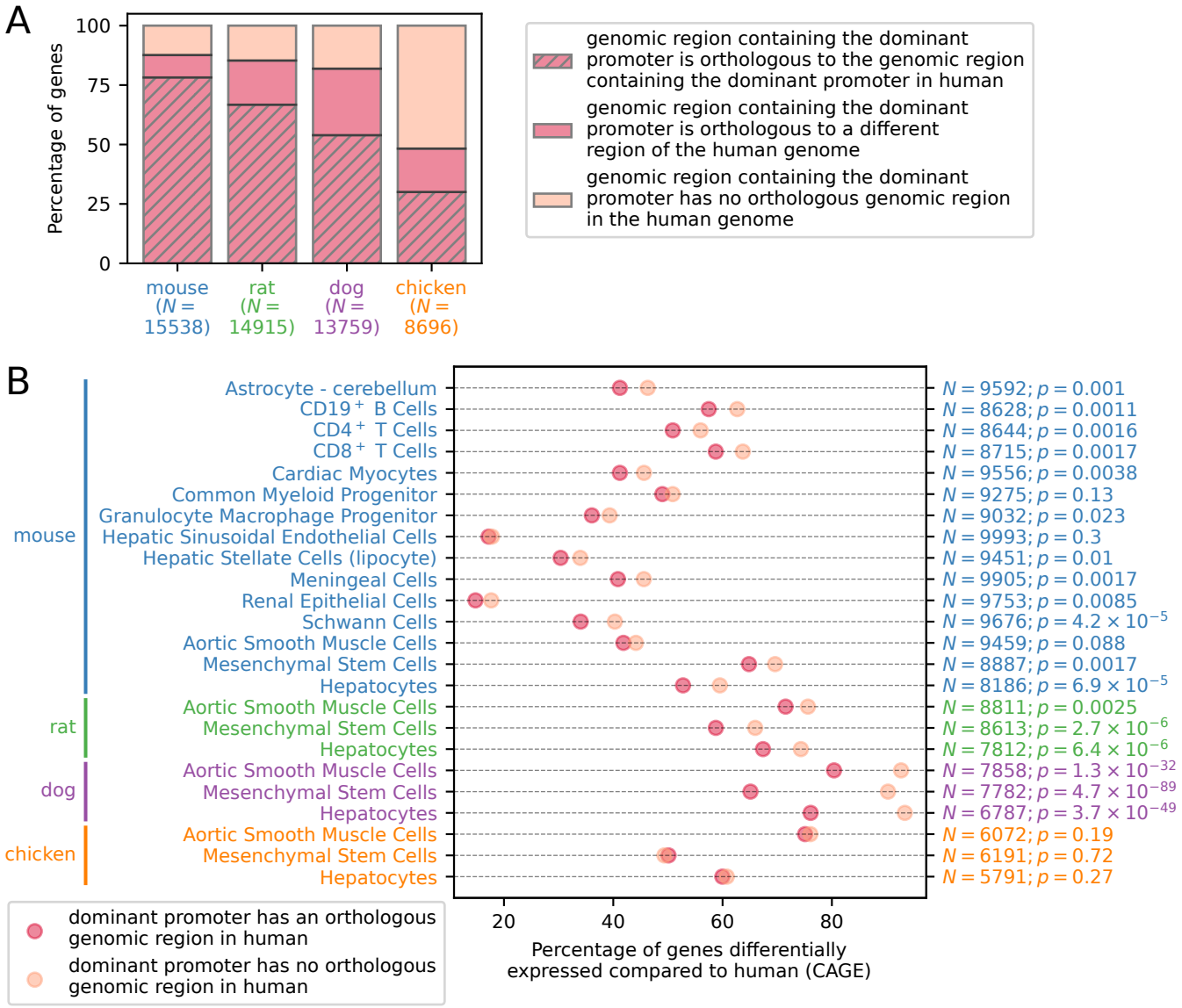
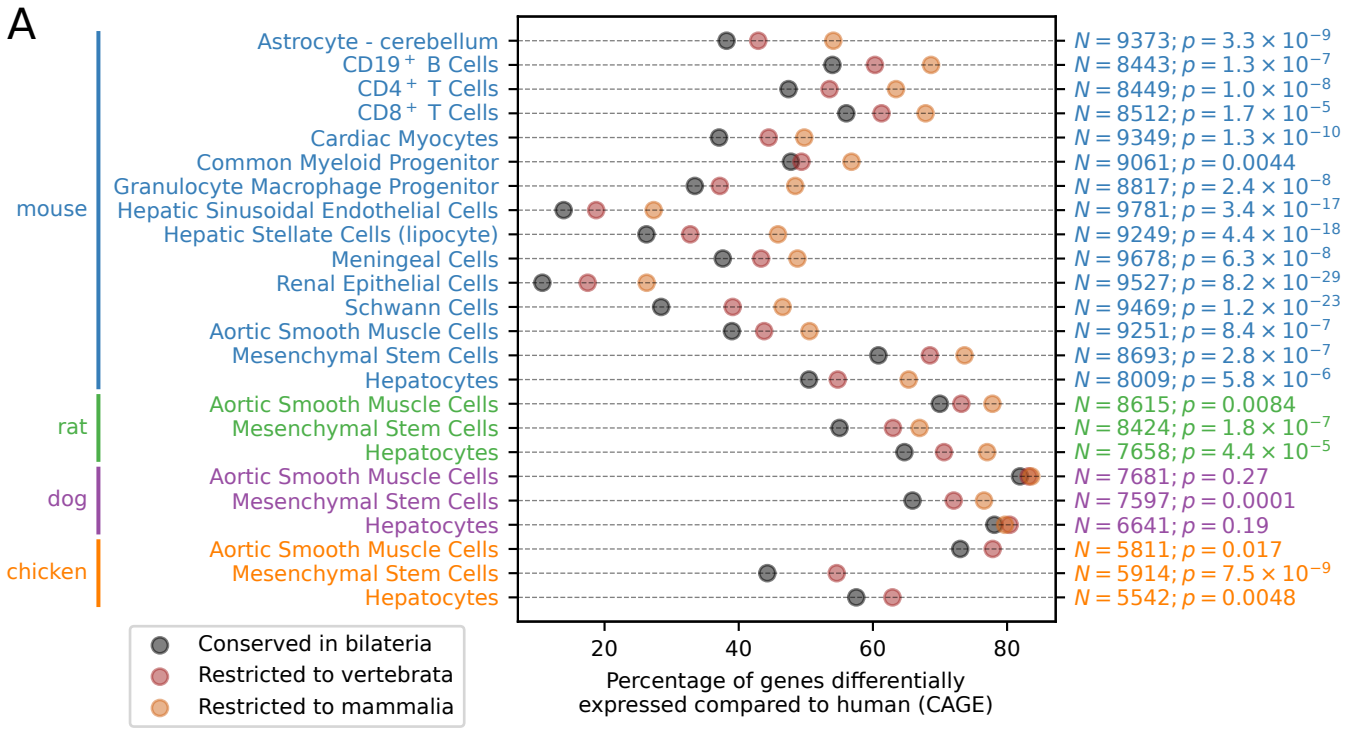

A
Percentage of genes
100
75
50
25
0
mouse (N = 15538)
rat (N = 14915)
dog (N = 13759)
chicken (N = 8696)
genomic region containing the dominant promoter is orthologous to the genomic region containing the dominant promoter in human
genomic region containing the dominant promoter is orthologous to a different region of the human genome
genomic region containing the dominant promoter has no orthologous genomic region in the human genome
B
mouse
rat
dog
chicken
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
N = 9592; p = 0.001
N = 8628; p = 0.0011
N = 8644; p = 0.0016
N = 8715; p = 0.0017
N = 9556; p = 0.0038
N = 9275; p = 0.13
N = 9032; p = 0.023
N = 9993; p = 0.3
N = 9451; p = 0.01
N = 9905; p = 0.0017
N = 9753; p = 0.0085
N = 9676; p = 4.2 × 10−5
N = 9459; p = 0.088
N = 8887; p = 0.0017
N = 8186; p = 6.9 × 10−5
N = 8811; p = 0.0025
N = 8613; p = 2.7 × 10−6
N = 7812; p = 6.4 × 10−6
N = 7858; p = 1.3 × 10−32
N = 7782; p = 4.7 × 10−89
N = 6787; p = 3.7 × 10−49
N = 6072; p = 0.19
N = 6191; p = 0.72
N = 5791; p = 0.27
20
40
60
80
Percentage of genes differentially expressed compared to human (CAGE)
dominant promoter has an orthologous genomic region in human
dominant promoter has no orthologous genomic region in human

Figure 4


A
mouse
rat
dog
chicken
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
N = 9373; p = 3.3 × 10−9
N = 8443; p = 1.3 × 10−7
N = 8449; p = 1.0 × 10−8
N = 8512; p = 1.7 × 10−5
N = 9349; p = 1.3 × 10−10
N = 9061; p = 0.0044
N = 8817; p = 2.4 × 10−8
N = 9781; p = 3.4 × 10−17
N = 9249; p = 4.4 × 10−18
N = 9678; p = 6.3 × 10−8
N = 9527; p = 8.2 × 10−29
N = 9469; p = 1.2 × 10−23
N = 9251; p = 8.4 × 10−7
N = 8693; p = 2.8 × 10−7
N = 8009; p = 5.8 × 10−6
N = 8615; p = 0.0084
N = 8424; p = 1.8 × 10−7
N = 7658; p = 4.4 × 10−5
N = 7681; p = 0.27
N = 7597; p = 0.0001
N = 6641; p = 0.19
N = 5811; p = 0.017
N = 5914; p = 7.5 × 10−9
N = 5542; p = 0.0048
Conserved in bilateria
Restricted to vertebrata
Restricted to mammalia
20
40
60
80
Percentage of genes differentially expressed compared to human (CAGE)

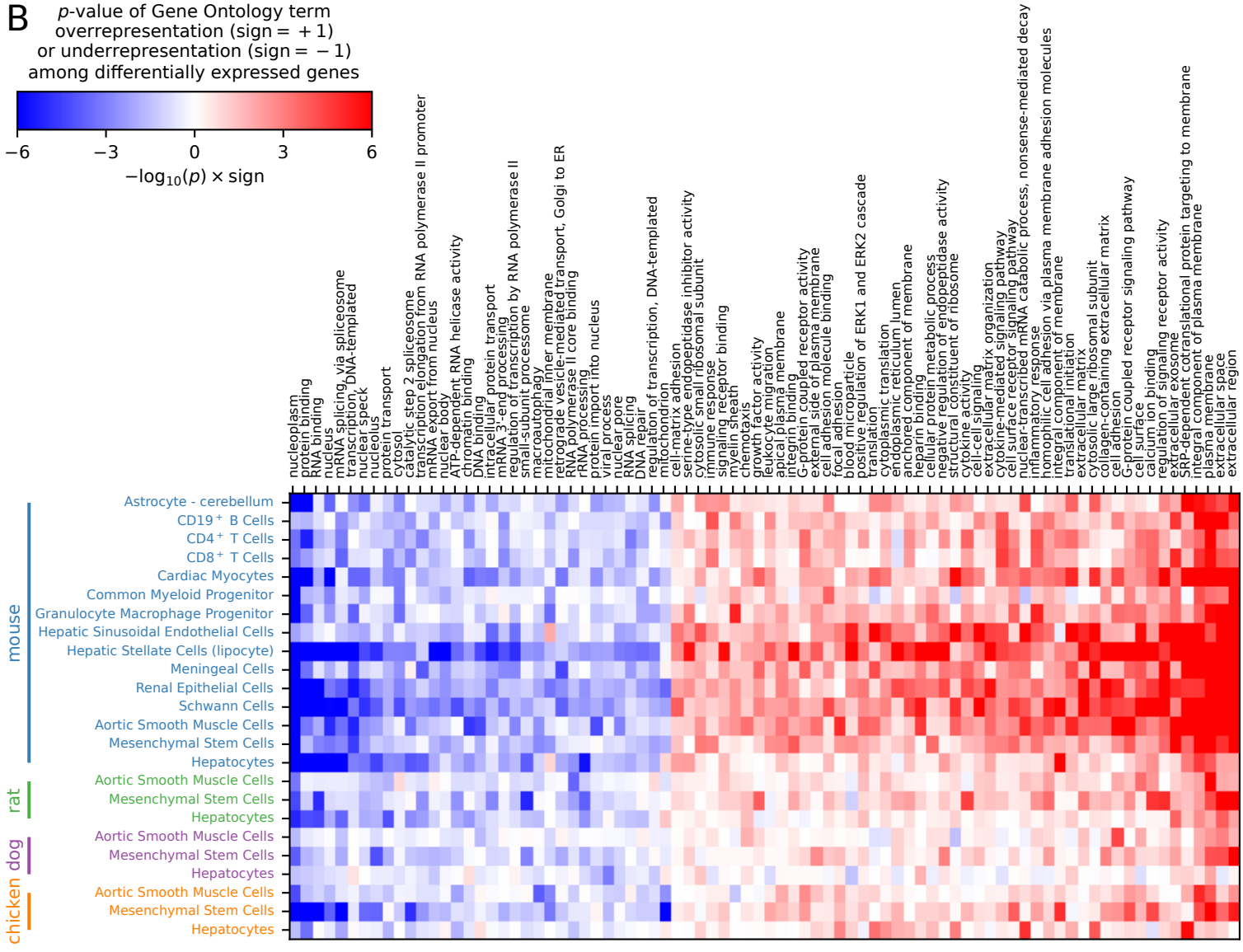

B
p-value of Gene Ontology term overrepresentation (sign = +1) or underrepresentation (sign = −1) among differentially expressed genes
−6
−3
0
3
6
−log10(p) × sign
nucleoplasm
protein binding
RNA binding
nucleus
mRNA splicing, via spliceosome
transcription, DNA-templated
nucleoplasm
nucleolus
protein transport
cytosol
catalytic step 2 spliceosome
transcription elongation from RNA polymerase II promoter
mRNA export from nucleus
nuclear body
ATP-dependent RNA helicase activity
chromatin binding
DNA binding
intracellular protein transport
mRNA 3'-end processing
regulation of transcription by RNA polymerase II
small-subunit processome
macroautophagy
mitochondrial inner membrane
retrograde vesicle-mediated transport, Golgi to ER
RNA polymerase II core binding
rRNA processing
protein import into nucleus
viral process
nuclear pore
RNA splicing
DNA repair
regulation of transcription, DNA-templated
mitochondrion
cell-matrix adhesion
serine-type endopeptidase inhibitor activity
cytosolic small ribosomal subunit
immune response
signaling receptor binding
myelin sheath
chemotaxis
growth factor activity
leukocyte migration
apical plasma membrane
integrin binding
G-protein coupled receptor activity
external side of plasma membrane
cell adhesion molecule binding
focal adhesion
blood microparticle
positive regulation of ERK1 and ERK2 cascade
translation
cytoplasmic translation
endoplasmic reticulum lumen
anchored component of membrane
heparin binding
cellular protein metabolic process
negative regulation of endopeptidase activity
structural constituent of ribosome
cytokine activity
cell-cell signaling
extracellular matrix organization
cytokine-mediated signaling pathway
cell surface receptor signaling pathway
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
inflammatory response
homophilic cell adhesion via plasma membrane adhesion molecules
integral component of membrane
translational initiation
extracellular matrix
cytosolic large ribosomal subunit
collagen-containing extracellular matrix
cell adhesion
G-protein coupled receptor signaling pathway
cell surface receptor
calcium ion binding
regulation of signaling receptor activity
extracellular exosome
SRP-dependent cotranslational protein targeting to membrane
integral component of plasma membrane
plasma membrane
extracellular space
extracellular region
mouse
rat
dog
chicken
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes

Figure 5

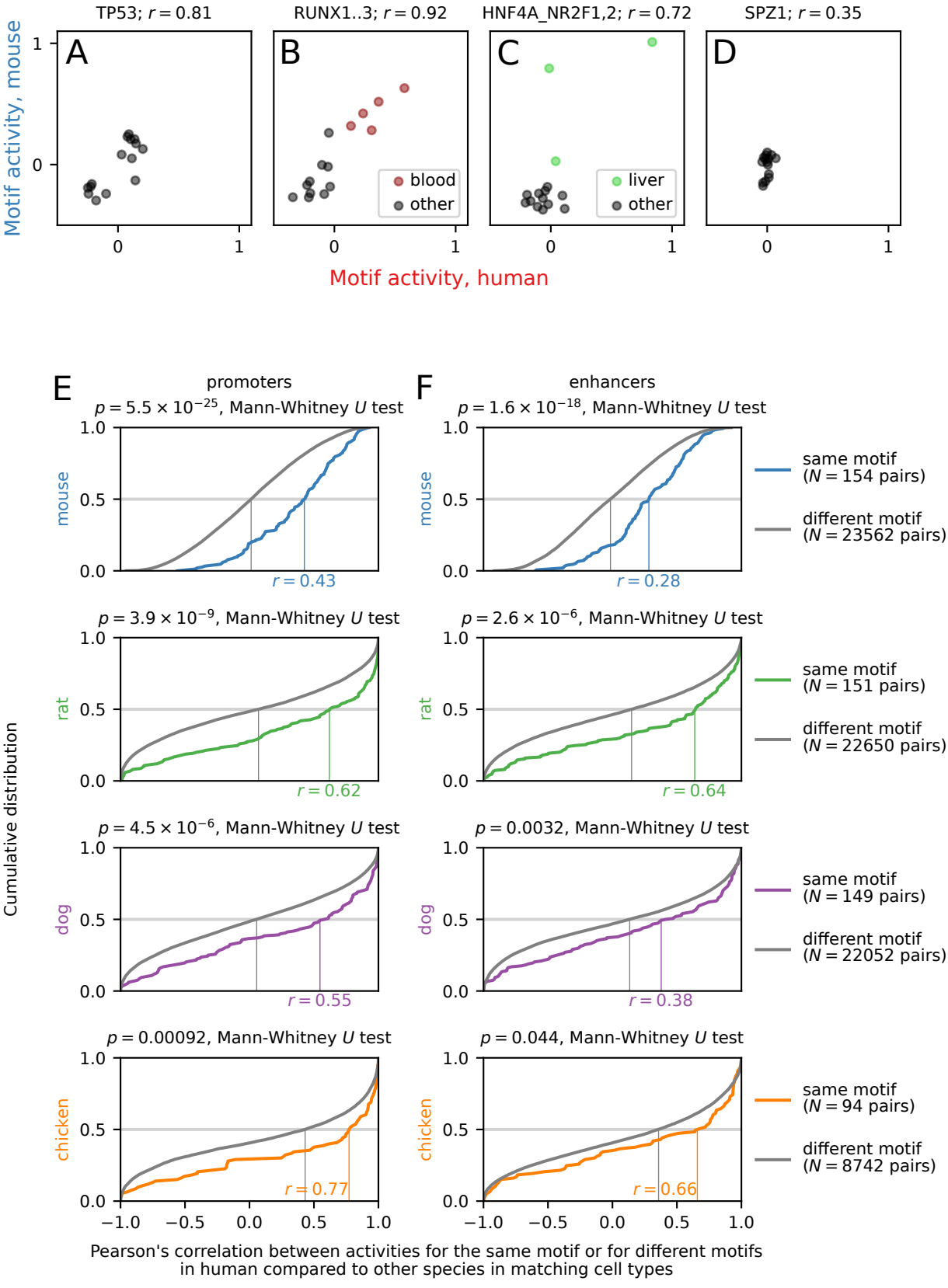
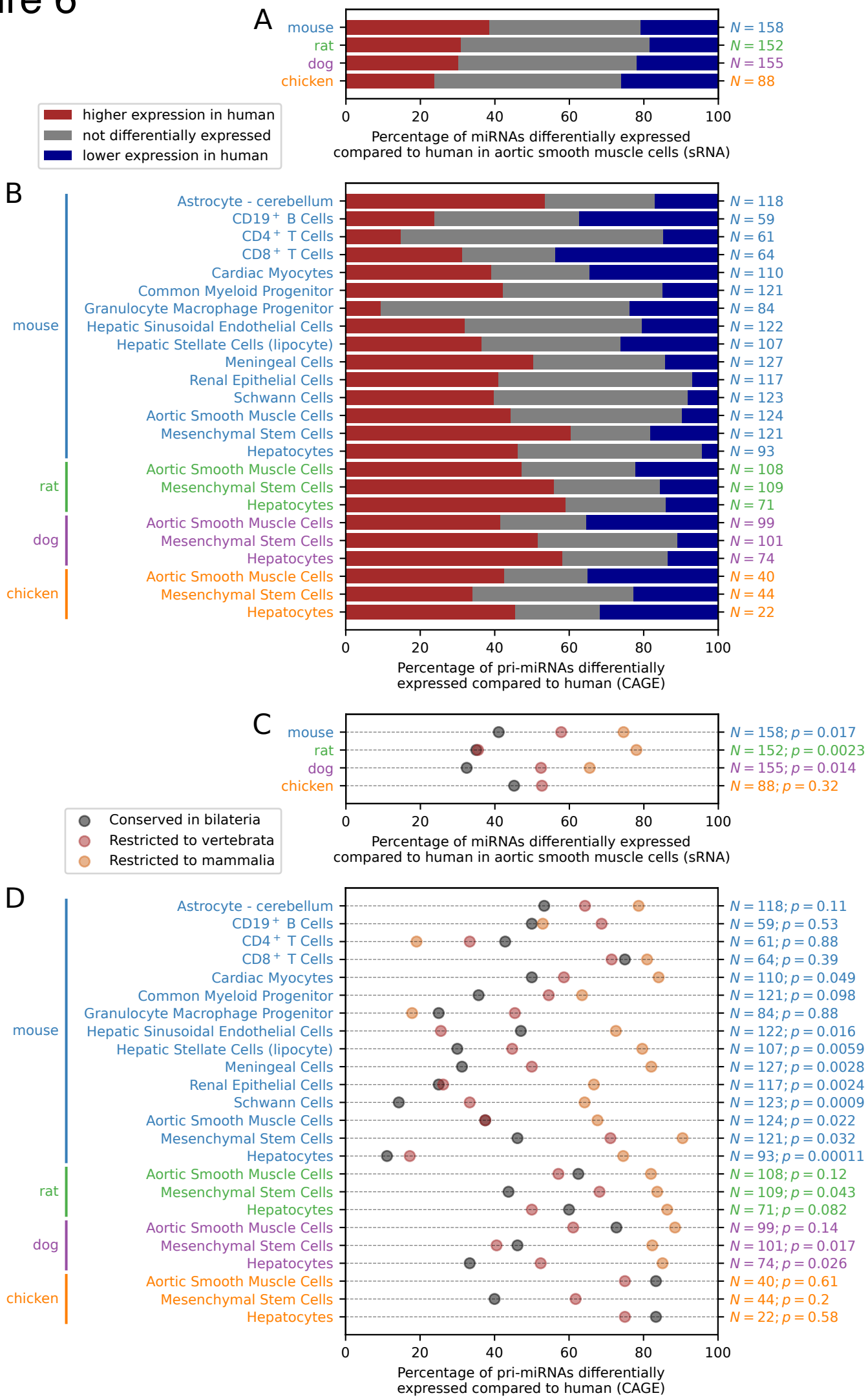

TP53; r = 0.81
RUNX1..3; r = 0.92
HNF4A_NR2F1,2; r = 0.72
SPZ1; r = 0.35
A
B
C
D
blood
other
liver
other
Motif activity, mouse
Motif activity, human
E
promoters
F
enhancers
p = 5.5 × 10−25, Mann-Whitney U test
p = 1.6 × 10−18, Mann-Whitney U test
p = 3.9 × 10−9, Mann-Whitney U test
p = 2.6 × 10−6, Mann-Whitney U test
p = 4.5 × 10−6, Mann-Whitney U test
p = 0.0032, Mann-Whitney U test
p = 0.00092, Mann-Whitney U test
p = 0.044, Mann-Whitney U test
mouse
rat
dog
chicken
r = 0.43
r = 0.28
r = 0.62
r = 0.64
r = 0.55
r = 0.38
r = 0.77
r = 0.66
same motif (N = 154 pairs)
different motif (N = 23562 pairs)
same motif (N = 151 pairs)
different motif (N = 22650 pairs)
same motif (N = 149 pairs)
different motif (N = 22052 pairs)
same motif (N = 94 pairs)
different motif (N = 8742 pairs)
Cumulative distribution
Pearson's correlation between activities for the same motif or for different motifs in human compared to other species in matching cell types

Figure 6


A
mouse
rat
dog
chicken
N = 158
N = 152
N = 155
N = 88
higher expression in human
not differentially expressed
lower expression in human
0
20
40
60
80
100
Percentage of miRNAs differentially expressed
compared to human in aortic smooth muscle cells (sRNA)
B
mouse
rat
dog
chicken
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
N = 118
N = 59
N = 61
N = 64
N = 110
N = 121
N = 84
N = 122
N = 107
N = 127
N = 117
N = 123
N = 124
N = 121
N = 93
N = 108
N = 109
N = 71
N = 99
N = 101
N = 74
N = 40
N = 44
N = 22
0
20
40
60
80
100
Percentage of pri-miRNAs differentially
expressed compared to human (CAGE)
C
mouse
rat
dog
chicken
N = 158; p = 0.017
N = 152; p = 0.0023
N = 155; p = 0.014
N = 88; p = 0.32
Conserved in bilateria
Restricted to vertebrata
Restricted to mammalia
0
20
40
60
80
100
Percentage of miRNAs differentially expressed
compared to human in aortic smooth muscle cells (sRNA)
D
mouse
rat
dog
chicken
Astrocyte - cerebellum
CD19+ B Cells
CD4+ T Cells
CD8+ T Cells
Cardiac Myocytes
Common Myeloid Progenitor
Granulocyte Macrophage Progenitor
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Meningeal Cells
Renal Epithelial Cells
Schwann Cells
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
Aortic Smooth Muscle Cells
Mesenchymal Stem Cells
Hepatocytes
N = 118; p = 0.11
N = 59; p = 0.53
N = 61; p = 0.88
N = 64; p = 0.39
N = 110; p = 0.049
N = 121; p = 0.098
N = 84; p = 0.88
N = 122; p = 0.016
N = 107; p = 0.0059
N = 127; p = 0.0028
N = 117; p = 0.0024
N = 123; p = 0.0009
N = 124; p = 0.022
N = 121; p = 0.032
N = 93; p = 0.00011
N = 108; p = 0.12
N = 109; p = 0.043
N = 71; p = 0.082
N = 99; p = 0.14
N = 101; p = 0.017
N = 74; p = 0.026
N = 40; p = 0.61
N = 44; p = 0.2
N = 22; p = 0.58
0
20
40
60
80
100
Percentage of pri-miRNAs differentially
expressed compared to human (CAGE)